Simulation with PARTS (phase-augmented research and training scenarios): a structure facilitating research and assessment in simulation

Schick, Carl J ; Weiss, Mona ; Kolbe, Michaela ; Marty, Adrian ; Dambach, Micha ; Knauth, Axel ; Spahn, Donat R ; Grote, Gudela ; Grande, Bastian

Abstract: INTRODUCTION: Assessment in simulation is gaining importance, as are scenario design methods increasing opportunity for assessment. We present our approach to improving measurement in complex scenarios using PARTS [Phase-Augmented Research and Training Scenarios], essentially separating cases into clearly delineated phases. METHODS: We created 7 PARTS with real-time rating instruments and tested these in 63 cases during 4 weeks of simulation. Reliability was tested by comparing real-time rating with postsimulation video-based rating using the same instrument. Validity was tested by comparing preintervention and postintervention total results, by examining the difference in improvement when focusing on the phase-specific results addressed by the intervention, and further explored by trying to demonstrate the discrete improvement expected from proficiency in the rare occurrence of leader inclusive behavior. RESULTS: Intraclass correlations [3,1] between real-time and postsimulation ratings were 0.951 (95% confidence interval [CI], 0.794-0.990), 1.00 (95% CI, –to–), 0.948 (95% CI, 0.783-0.989), and 0.995 (95% CI, 0.977-0.999) for 3 phase-specific scores and total scenario score, respectively. Paired t tests of prelecture-postlecture performance showed an improvement of 14.26% (bias-corrected and accelerated bootstrap [BCa] 95% CI, 4.71-23.82; P = 0.009) for total performance but of 28.57% (BCa 95% CI, 13.84-43.30; P = 0.002) for performance in the respective phase. The correlation of total scenario performance with leader inclusiveness was not significant (rs = 0.228; BCa 95% CI. -0.082 to 0.520; P = 0.119) but significant for specific phase performance (rs = 0.392; BCa 95% CI, 0.118-0.632; P = 0.006). CONCLUSIONS: The PARTS allowed for improved reliability and validity of measurements in complex scenarios.

DOI: https://doi.org/10.1097/SIH.0000000000000085

Simulation With PARTS (Phase-Augmented Research and Training Scenarios)

A Structure Facilitating Research and Assessment in Simulation

Carl J. Schick;
Mona Weiss, PhD;
Michaela Kolbe, PhD;
Adrian Marty, MD;
Micha Dambach, MD;
Axel Knauth, MD;
Donat R. Spahn, MD;
Gudela Grote, PhD;
Bastian Grande, MD

Introduction: Assessment in simulation is gaining importance, as are scenario design methods increasing opportunity for assessment. We present our approach to improving measurement in complex scenarios using PARTS [Phase-Augmented Research and Training Scenarios], essentially separating cases into clearly delineated phases.
Methods: We created 7 PARTS with real-time rating instruments and tested these in 63 cases during 4 weeks of simulation. Reliability was tested by comparing real-time rating with postsimulation video-based rating using the same instrument. Validity was tested by comparing preintervention and postintervention total results, by examining the difference in improvement when focusing on the phase-specific results addressed by the intervention, and further explored by trying to demonstrate the discrete improvement expected from proficiency in the rare occurrence of leader inclusive behavior.
Results: Intraclass correlations [3,1] between real-time and postsimulation ratings were 0.951 (95% confidence interval [CI], 0.794–0.990), 1.00 (95% CI, — to —), 0.948 (95% CI, 0.783–0.989), and 0.995 (95% CI, 0.977–0.999) for 3 phase-specific scores and total scenario score, respectively. Paired t tests of prelecture-postlecture performance showed an improvement of 14.26% (bias-corrected and accelerated bootstrap [BCa] 95% CI, 4.71–23.82; P = 0.009) for total performance but of 28.57% (BCa 95% CI, 13.84–43.30; P = 0.002) for performance in the respective phase. The correlation of total scenario performance with leader inclusiveness was not significant (r_s = 0.228; BCa 95% CI. −0.082 to 0.520; P = 0.119) but significant for specific phase performance (r_s = 0.392; BCa 95% CI, 0.118–0.632; P = 0.006).
Conclusions: The PARTS allowed for improved reliability and validity of measurements in complex scenarios.
(*Sim Healthcare* 10:178–187, 2015)

Key Words: Crisis resource management, Simulation, Training, Scenario design, Assessment, Rating, Complex cases, Visual rating tool, Phase separation, Debriefing, Facilitation, Feedback, Measurement, Anesthesia.

As an educational instrument, simulation in health care regularly involves providing feedback as an important element promoting learning.[1] Debriefing, a feedback process especially suited for experiential learning, in which a facilitator helps in bridging the gap between the experience and the "making sense of it,"[2] will often be used in simulation-based training (SBT) involving team training, crew resource management skills, and multidisciplinary interactions. Common approaches to this

From the Institute of Anesthesiology (C.J.S., A.M., M.D., A.K., D.R.S., B.G.), University Hospital Zürich; and Department of Management, Technology, and Economics Organization (M.W., M.K., G.G.), Work and Technology Group, ETH Zürich, Zürich, Switzerland.

Reprints: Carl J. Schick, Institute of Anesthesiology, University Hospital Zürich, Rämistrasse 100, 8091 Zürich, Switzerland (e-mail: carl.schick@usz.ch).
Dr Spahn's academic department is receiving grant support from the Swiss National Science Foundation, Berne, Switzerland (grant numbers: 33CM30_124117 and 406440–131268); the Swiss Society of Anesthesiology and Reanimation (SGAR), Berne, Switzerland (no grant numbers are attributed); the Swiss Foundation for Anesthesia Research, Zurich, Switzerland (no grant numbers are attributed); Bundesprogramm Chancengleichheit, Berne, Switzerland (no grant numbers are attributed); CSL Behring, Berne, Switzerland (no grant numbers are attributed); and Vifor SA, Villars-sur-Glâne, Switzerland (no grant numbers are attributed).

Dr Spahn was the chairman of the ABC Faculty and is the cochairman of the ABC-Trauma Faculty, which both are managed by Physicians World Europe GmbH, Mannheim, Germany, and sponsored by unrestricted educational grants from Novo Nordisk Health Care AG, Zurich, Switzerland; CSL Behring GmbH, Marburg, Germany; and LFB Biomédicaments, Courtaboeuf Cedex, France.

In the past 5 years, Dr Spahn has received honoraria or travel support for consulting or lecturing from the following companies: Abbott AG, Baar, Switzerland; AMGEN GmbH, Munich, Germany; AstraZeneca AG, Zug, Switzerland; Bayer (Schweiz) AG, Zürich, Switzerland; Baxter AG, Volketswil, Switzerland; Baxter S.p.A., Roma, Italy; B. Braun Melsungen AG, Melsungen, Germany; Boehringer Ingelheim (Schweiz) GmbH, Basel, Switzerland; Bristol-Myers-Squibb, Rueil-Malmaison Cedex, France and Baar, Switzerland; CSL Behring GmbH, Hattersheim am Main, Germany and Berne, Switzerland; Curacyte AG, Munich, Germany; Ethicon Biosurgery, Sommerville, NJ; Fresenius SE, Bad Homburg v.d.H., Germany; Galenica AG, Bern, Switzerland (including Vifor SA, Villars-sur-Glâne, Switzerland); GlaxoSmithKline GmbH & Co. KG, Hamburg, Germany; Janssen-Cilag AG, Baar, Switzerland; Janssen-Cilag EMEA, Beerse, Belgium; Merck Sharp & Dohme-Chibret AG, Opfikon-Glattbrugg, Switzerland; Novo Nordisk A/S, Bagsvärd, Denmark; Octapharma AG, Lachen, Switzerland; Organon AG, Pfäffikon/SZ, Switzerland; Oxygen Biotherapeutics, Costa Mesa, CA; Pentapharm GmbH (now tem Innovations GmbH), Munich, Germany; ratiopharm Arzneimittel Vertriebs-GmbH, Vienna, Austria; Roche Pharma (Schweiz) AG, Reinach, Switzerland; Schering-Plough International, Inc, Kenilworth, NJ; Vifor Pharma Deutschland GmbH, Munich, Germany; Vifor Pharma Österreich GmbH, Vienna, Austria; and Vifor (International) AG, St. Gallen, Switzerland.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.simulationinhealthcare.com).

DOI: 10.1097/SIH.0000000000000085

structured process, especially the established framework for debriefing with good judgment,[3] consider salient performance gaps related to predetermined objectives. Hence, although SBT does not rely on accurate performance measurement, precise assessment of simulation performance seems to improve the scope of educational opportunity and is required by many forms of feedback. In addition, simulation in health care is increasingly being used as a method for outcome measurement, such as high-stakes testing, recertification, and translational research,[4–9] that is, the transfer of results from bench to bedside.[10] This application of simulation in health care depends on reliable measurement of performance.

Performance in complex scenarios depicting critical incidents is difficult to assess and particularly so in real time possibly because of the interweaving nature of storylines, unexpected learner actions, and their duration. In an attempt to improve measurement in these scenarios, we developed and tested a scenario and rating design method integrating elements previously described in educational and simulation research.[11–21] The resulting structure, Phase-Augmented Research and Training Scenarios (PARTS), allows for separating complex cases into clearly delineated phases with single critical events. The goal of PARTS is to provide a scenario development process, allowing for specifically targeting learning objectives and their real-time measurement.

In this article, we will describe the rationale for standardized scripting of simulated cases, followed by details on the development of PARTS, and present PARTS with the results of reliability and validity tests. With this scenario design method, we aimed to foster SBT and research, eventually leading to improvements in clinical performance.

BACKGROUND

During the last 4 years, our simulation center performed approximately 100 days of crisis resource management (CRM) training using high-fidelity patient simulators for staff from our anesthesiology department. Cases were vaguely based on reported critical incidents and immediately followed by debriefing facilitated by a senior anesthetist and a senior psychologist using TeamGAINS,[22] a hybrid concept for debriefing. Trained independent raters attempted to examine performance using video recordings and ordinal rating lists, at time including close to 100 items. This approach was found to have substantial disadvantages as follows:

1. Comprehensive rating tools designed for in-depth video-based evaluation could not be used for real-time rating to provide debriefers with objective performance measurement.
2. Lacking focused measurement, the effect of education interventions only treating a specific part of the case on repeated (pre-post) participation in the same scenario could not be isolated from the overall improvement expected because of repetitive exposure.
3. Because of the lack of standardization, comparing performance in similar situations across different scenarios was difficult.
4. Measurement of discrete effects expected from proficiency in specific competences affecting only a small part of the scenario performance was difficult because of the diluent effect of overall performance scores.

In light of these drawbacks, we attempted to improve our approach to the scenario design. The goals of our study were to develop a method for standardized scenario development and to test this method for reliability and validity. With respect to reliability, we assumed that the PARTS scenario flowchart would allow for real-time rating providing similar results to postsimulation, video-based rating (hypothesis 1). Concerning validity, we expected that PARTS would allow for detecting differences in performance after an educational intervention, that is, reveal overall higher postintervention scores than preintervention scores (hypothesis 2) and even higher postintervention scores focusing on the part of the scenario specifically addressed by the intervention (hypothesis 3). To further explore the validity of PARTS, we tested whether PARTS performance ratings would reflect variations in team coordination, specifically in leader inclusive behavior. Leader inclusiveness is a behavior in which the leader explicitly invites team members to share their opinions and suggestions and appreciates their contributions.[23] It is considered important for establishing psychological safety, allowing team members to engage in the team process, and particularly beneficial in situations of increased task complexity. Assuming that leader inclusiveness is generally rare, we expected that total scenario rating would not allow for detecting associations with performance (hypothesis 4), whereas specific rating focused on phases with high task complexity would (hypothesis 5).

METHODS

Development of PARTS

In a first step, we reviewed the literature on assessment and scenario design.[1,3,9,16–18,20,22,24–72] As we will describe, we found established and tested scenario design techniques, which we incorporated into PARTS.

A prominent finding of the literature was the separation of scenarios into phases, such as "before and after declaration of an emergency"[11] or "preparation, pre-intubation and intubation."[12] Although phase separation can be subjective, different tasks usually have different coordination requirements,[13] for example, information management during diagnostic phases versus direct leadership during resuscitation phases.[14] Accordingly, separating cases into phases based on the coordination requirements (eg, gathering information from the team) and respective objectives (finding the cause of cardiac arrest) could lead to improved task-specific measurements. In addition, this delineation could help in comparing similar phases across different scenarios.

Most critical incidents evolve in a similar pattern, consisting of the following:

1. A preliminary phase, in which the patient is initially stable, but anticipation and meticulous preparation might mitigate the effects of an ensuing crisis;
2. An emergency phase with a deteriorating patient and increased task complexity, profiting from shared leadership,[73] increased team coordination, and communication to achieve patient stabilization and the establishment of a diagnosis; and

3. A management phase involving the treatment of a found diagnosis, requiring a clear lead, delegation, and more procedural task performance.

In a second stage, our instructor team of 5 anesthetists and 2 psychologists (A.K., A.M., B.G., C.J.S., M.D., M.K., M.W.) selected reported critical incidents suitable for CRM training. We anticipated that the use of real cases would contribute to content validity.[38] Our aim was to identify the 3 aforementioned phases in these cases and the single main critical event matching each phase, such as a cardiac arrest caused by unapparent pneumothorax requiring decompression in the emergency phase and postresuscitation care in the management phase. Where the original case did not provide material for all 3 phases (ie, preliminary, emergency, and management), we created the missing critical event in accordance with the learning objective. For example, designing a scenario for training a complete handover based on a reported case without a preliminary phase, we had the paramedic attempting to leave before providing all necessary information, thus creating the incomplete handover as the critical event in the preliminary phase. This formed the basis of each scenario template as illustrated in Figure 1.

To clearly separate phases, we decided on observable markers of phase transition, allowed times per phase, and noted these on the scenario template. Often, phase transitions were instructor-controllable events such as the onset of cardiac arrest at a specific time. Transitions relying on participant actions, such as the statement of the correct diagnosis, required a backup, that is, a lifesaver,[16] which could be used to nudge stalled or fixated[74] participants on to the next phase should the time limit be reached. For example, an instructor acting as a surgeon might announce the myocardial infarction visible on the patient monitor.

Should the information not be clear to the participants or unexpected problems such as unnoticed esophageal intubation arise, we would resort to an instructor entering the simulation room and clarifying, by stating "Time out—(clarification). Please continue the scenario accordingly."

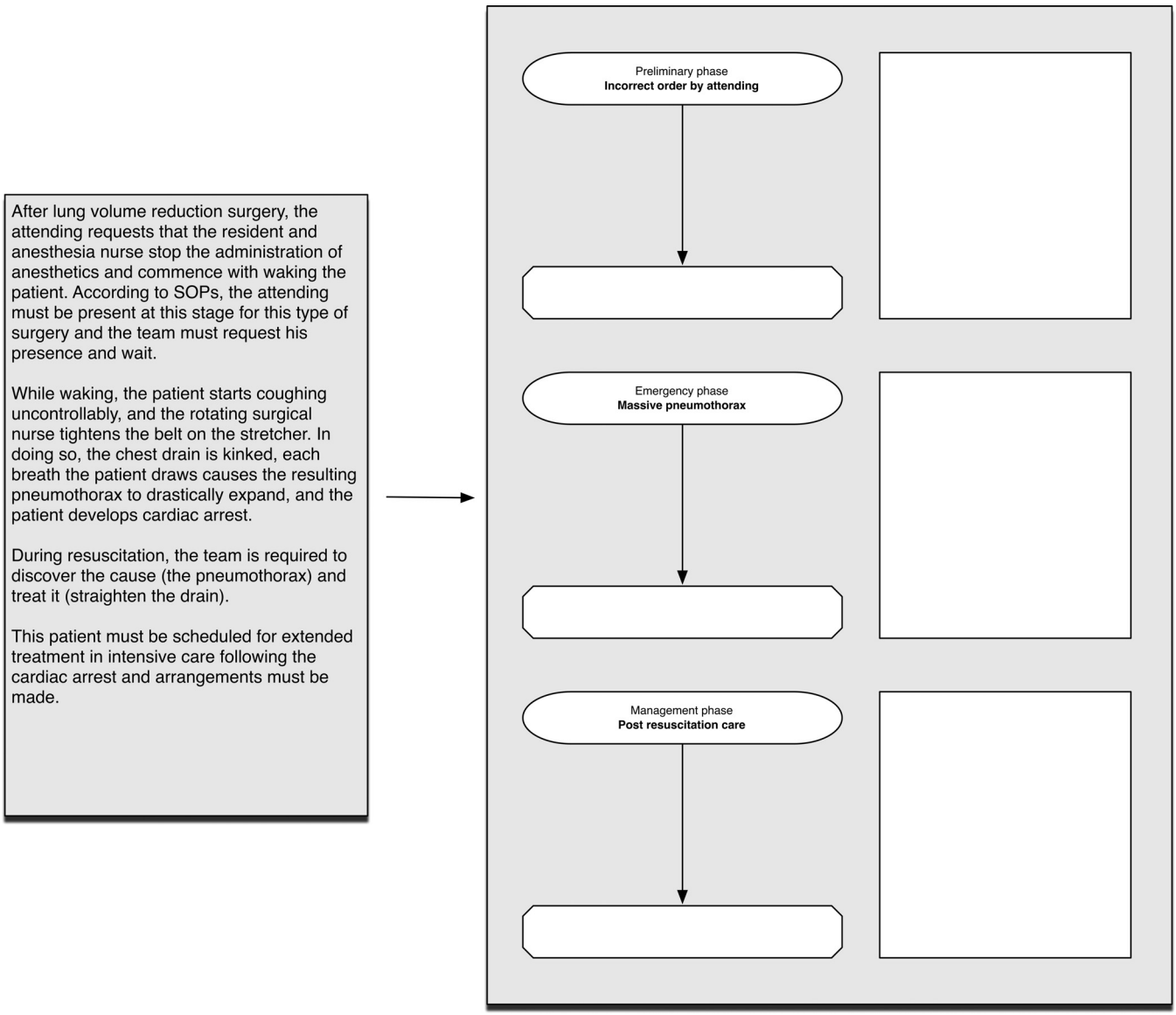

FIGURE 1. An example of a reported critical incident (left) analyzed for main critical events per phase, which are added to the respective scenario template (right) showing the phases and the selected critical events.

Although this could potentially disrupt scenario flow and participant performance, recent results suggest that this is not necessarily the case,[75] and we similarly hoped this would allow the next phase to commence as planned with a "clean slate," regardless of current performance and actions. We used this backup only once, when a team became fixated on a pneumothorax instead of a massive hemorrhage causing cardiac arrest caused by a malfunctioning loudspeaker in the mannequin's chest. Although the team had to reset and reorganize itself after the time-out, the next phase (treatment of massive hemorrhage) commenced and ran as expected.

In a third step, we developed a process for evaluation of performance during the phases. Considering the literature, we found that many comprehensive approaches to scenario design such as SMARTER [the Simulation Module for Assessment of Resident Targeted Event Responses] for emergency medicine residents[17,18] and TARGETs [Targeted Acceptable Responses to Generated Events or Tasks], a methodology for improving measurement of military team performance,[19] are adaptations of an event-based approach to training (EBAT),[20] which drives scenario development around a defined single critical event requiring specific learner actions. In turn, these can be directly transcribed to a rating instrument, scoring performance by counting completed items on a checklist. We used EBAT for each phase within each scenario as shown in Figure 2.

We used the Delphi technique[15] to derive the required learner actions during critical events by expert consensus, increasing face and content validity of the selection. Each scenario designer (A.K., A.M., B.G., C.J.S., M.D., M.K., M.W.) noted all required learner actions they felt were necessary to perfectly solve each critical event. Results were compared, and those suggested by at least 80% of designers were included in the final list.

To address the problem of rating unobservable events, results of the Delphi rounds were reviewed for visibility in the scenario (ie, whether they could observed). In some instances, we defined surrogate markers acceptable despite their reduced specificity for the required learner action. For example, merely considering a pneumothorax is not observable, but auscultation and/or requesting information on the result is and would be considered an indicator of searching for a pneumothorax.

To allow for expert participants legitimately skipping steps, we used hierarchical task analysis[14] to examine required learner actions for subtasks, which need not necessarily be performed but which may prove helpful to the less experienced clinician. For example, ordering laboratory results, auscultation, monitoring the heart rhythm, administering fluids, and verifying oxygenation (representing the discovered subtasks) can help in discovering the cause of cardiac arrest (the required learner action) but are not necessary if the cause is otherwise discovered.

Figure 3 illustrates an extract of the resulting rating instrument and demonstrates scoring based on 2 different examples.

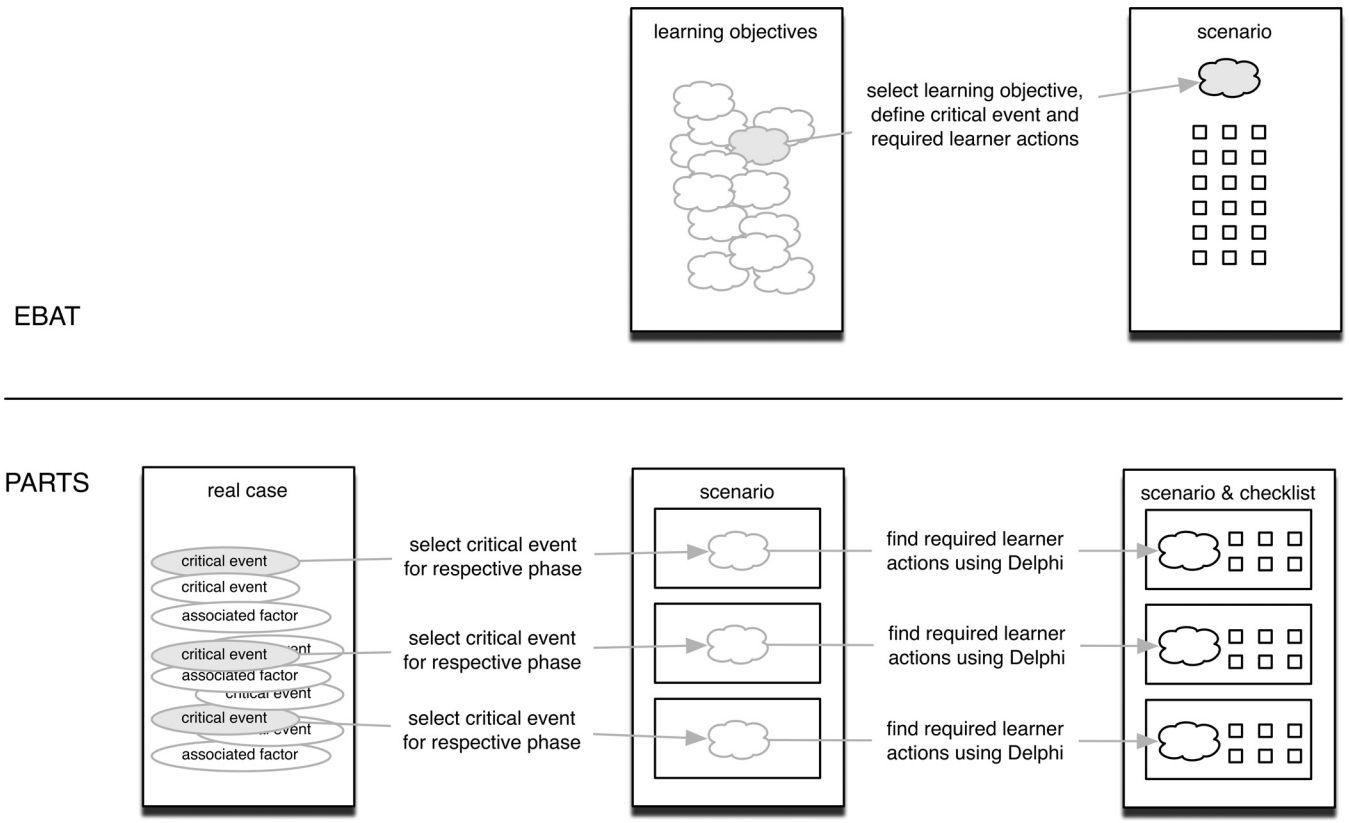

FIGURE 2. Comparison of the scenario design workflow with traditional EBAT scenario design and PARTS design, which uses the EBAT for each phase. Extending EBAT scenario design around 1 critical event requiring learner actions, PARTS enables designers to chain critical events, each with their respective learner actions, while clearly separating these for individual rating.

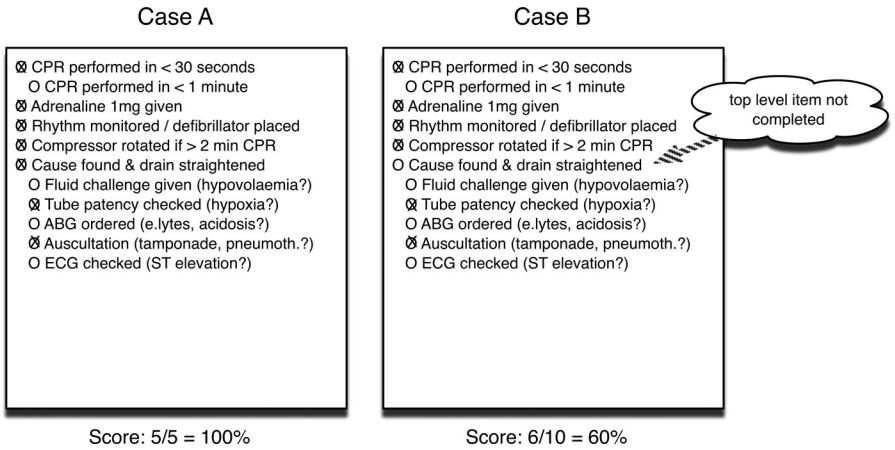

FIGURE 3. Extract of the rating instrument for the emergency phase (resuscitation and search for the cause) in the pneumothorax case showing the results of the Delphi process and hierarchical task analysis and comparing 2 possible team ratings. In case A, an expert skipping some subtasks, immediately discovering the cause and straightening the drain in light of the clinical expertise or previous experience, is not penalized. The completed top-level items (5/5) give a score of 100%. Conversely, if the cause of cardiac arrest is not discovered (case B), relevant subtasks such as the search for cardiac tamponade or pneumothorax by auscultation are considered an important achievement. The first 4 items are completed (4/4), whereas the last item was not; hence, its 5 subtasks are also included in the rating, of which 2 (2/6) are completed, giving a total phase score of 6/10 or 60%.

Testing PARTS for Reliability and Validity

We tested PARTS for reliability and validity during SBTs for anesthesia staff. This prospective study was exempt from institutional review board approval by the ethics committee of Zurich, Switzerland. Written consent was obtained from all study participants.

Sample and Procedure

The study was conducted during 2 blocks in January and March 2013, which were structured as illustrated in Figure 4.

During 10 days in January and 9 days in March 2013, 117 individual members of anesthesia staff (11 attendings, 57 residents, and 49 nurses) participated in SBTs. On each training day, 1 attending, 3 residents, and 3 nurses were present (9 attendings and 8 nurses participated in both the January and March rounds). Each day, residents and nurses were assigned to 1 of 3 groups at random, with the attending available to help each group whenever called. Each group participated in 1 scenario; on days where time allowed for 4 scenarios, 1 voluntary group participated in another case. The other groups remained in the debriefing room watching via video transmission and participated in the debriefings.

In the January round, we presented scenarios 1 to 5 in varying order because of availability of equipment and to reduce sequence bias. In March, we used scenario 6 before and after scenario 7 to perform pre-comparison - post-comparison (Document, Supplemental Digital Content 1, http://links.lww.com/SIH/A208, for the implementation schedule and scenario flowcharts for cases 1–7). Each scenario lasted approximately 20 minutes with debriefings taking approximately an hour.

The scenarios were programmed using the Laerdal SimMan3G scenario editor software and only needed adapting when unexpected actions, such as an unexpected drug administration, occurred. The instructor controlling the mannequin (C.J.S.) was involved in both scenario programming and design and hence felt comfortable with controlling the simulation and performing real-time rating simultaneously—no further instructors performed real-time ratings. For each scenario, percentages of item completion were calculated for each phase as well as for the complete scenario, resulting in 4 ratings per case (Fig. 5).

Testing PARTS for Reliability

To test whether the PARTS scenario flowchart would allow for real-time rating and provide similar results to postsimulation, video-based rating, we compared these ratings for 8 scenarios randomly selected from the 63 cases.

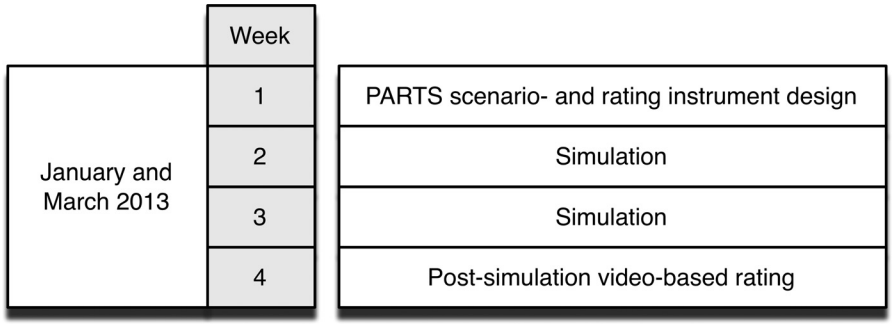

FIGURE 4. Timeline of study blocks in January and March 2013. A week of scenario design was followed by 2 weeks of simulation, with the fourth week being used for postsimulation video-based rating.

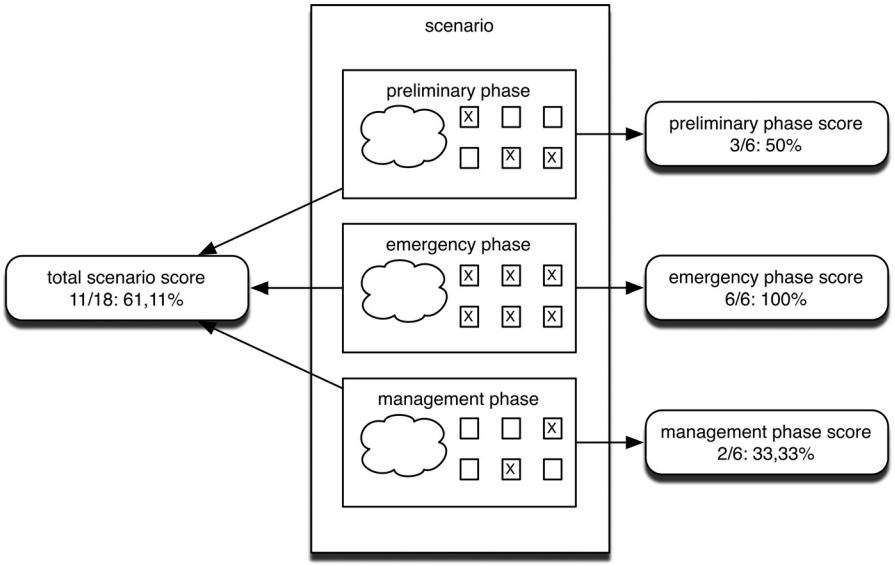

FIGURE 5. Diagram of scores obtained for each scenario. Total scores were obtained by calculating the percentage of all completed items, irrespective of phase.

This amount was decided upon based on the time the independent raters could invest. Video recordings of the scenarios were used by 2 independent anesthetists, neither involved in the scenario design process nor trained in the use of the rating instrument, to each rate 4 of the 8 selected cases using the identical scenario flowchart.

Testing PARTS for Validity

To analyze whether PARTS would allow for valid performance assessment, we applied 2 approaches.

To test whether PARTS would allow for detecting differences in performance after an educational intervention, we applied a pretest intervention–posttest design. During each of 9 days in the second study block, learners participated in scenario 6, which was followed by a debriefing focused on the treatment of massive hemorrhage—the critical event of the management phase. Later that day, learners participated in this scenario with an additional patient as a distractor. We compared the total as well as the management phase performance ratings before and after the intervention.

Second, to test whether PARTS performance ratings would reflect variations in team coordination, specifically in leader inclusive behavior[23] presumed to be beneficial during the emergency phases of scenarios 1, 2, and 6 (with a rapidly deteriorating patient because of unknown reasons), the raters performing postsimulation video-based rating also examined these 48 phases. Using the videos, they counted the number of statements inviting or asking for ideas, opinions, or help from the team, issued by the physician with the highest hierarchical hospital position participating in the case. Subsequently, we analyzed the relation between the number of resulting leader inclusive statements and the respective total as well as emergency phase performance ratings.

Statistical Analysis

Data were analyzed using SPSS 21.0. Hypothesis 1 was tested using intraclass correlation coefficients (ICCs [3,1]) between real-time and postsimulation ratings. Hypotheses 2 and 3 were tested using a paired-samples *t* test comparing premanagement and postmanagement phase and total scenario scores for scenario 6. Hypotheses 4 and 5 were examined measuring correlations of leader inclusiveness with standardized emergency phase and total scenario scores in scenarios 1, 2, and 6.

RESULTS

PARTS Scenario Flowchart

The final PARTS scenario flowchart is shown in Figure 6.

Reliability of PARTS

In hypothesis 1, we predicted that the PARTS scenario flowchart would allow for real-time rating providing similar results to postsimulation, video-based rating. Intraclass correlation showed high agreement between these ratings throughout, supporting hypothesis 1 (Table 1).

Validity of PARTS

In hypotheses 2 and 3, we postulated that PARTS would allow for detecting differences in performance after an educational intervention, that is, to reveal higher total postintervention than preintervention scores (hypothesis 2), and even higher postintervention scores when only taking the respective phase into account (hypothesis 3). Preliminary analysis showed normal distribution of mean differences with no outliers on visual inspection. Results of the paired *t* test demonstrated that the total score shows a mean pretest-posttest improvement of 14.26% ($d = 1.56$, $P = 0.009$), whereas management phase-specific score shows a mean pretest-posttest improvement twice as large (28.57%; $d = 2.22$, $P = 0.002$), supporting both hypotheses (Table 2).

Assuming that leader inclusiveness is generally rare but particularly important with increased task complexity, in hypothesis 4, we postulated that total performance rating would not allow for detecting associations with leader inclusive behavior across the complete scenario. In hypothesis

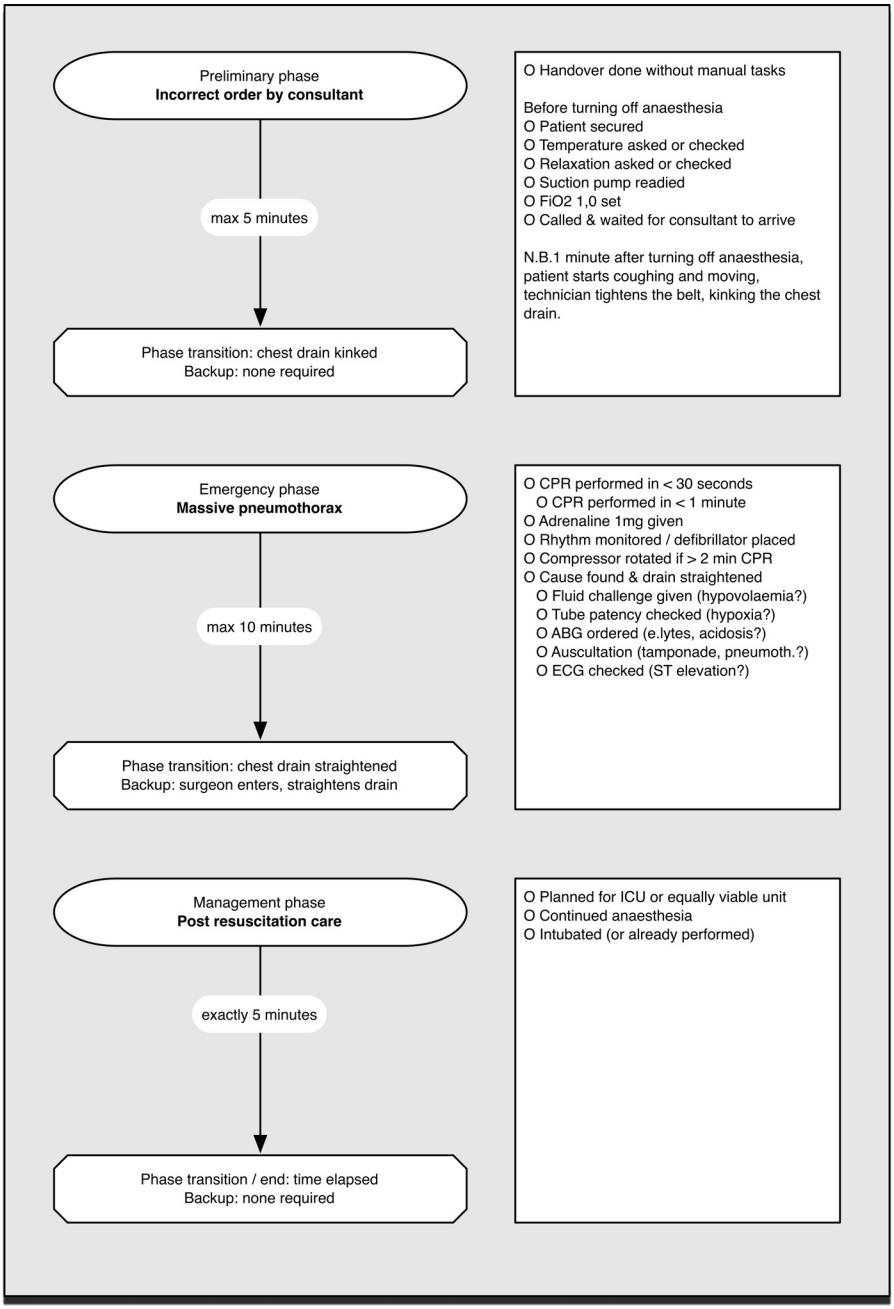

FIGURE 6. Completed scenario flowchart depicting the separate phases, critical events, time limits per phase, phase transitions and backups, together with the required learner actions with respective check boxes to be used for real-time rating.

5, we assumed that by using PARTS to attain specific scores for the phase expected to require this particular leadership behavior, the effect on performance might become visible, albeit with the small overall effect expected of one of many variables contributing to team performance.

Correlations of leader inclusiveness with emergency phase and total scores for scenarios 1, 2, and 6 were examined. Preliminary analysis of these results showed linear relationships with no outliers on visual inspection. However, leader inclusiveness was moderately right skewed; therefore, a Spearman rank-order correlation was performed. Examining the total scenario scores of all combined cases, we found a low and statistically insignificant correlation of performance with leader inclusiveness (r_s = 0.228, P = 0.119), supporting hypothesis 4. When examining the emergency phase scores, however, we found a higher and statistically significant correlation of the phase-specific performance with

TABLE 1. Intraclass Correlations [3,1] for Real-time and Postsimulation Ratings of 8 Randomly Selected Scenarios

Score	ICC*	95% CI	*P*
Preliminary phase score	0.951	0.794 to 0.990	0.000
Emergency phase score	1.00	— to —	0.000
Management phase score	0.948	0.783 to 0.989	0.000
Total scenario score	0.995	0.977 to 0.999	0.000

*Two-way mixed-effects model for absolute agreement.
Interrater reliability is examined for individual phase scores and for total scenario scores obtained with the same rating instrument but different untrained raters.

TABLE 2. Paired t Test Results of Premeasurement-Postmeasurement for Performance in Scenario 6, Which Was Presented Before and After an Educations Intervention Specifically Addressing the Management Phase

Outcome	Pretest		Posttest		n	Mean Difference	BCa 95% CI*	Effect Size *d*	*P*
	Mean	SD	Mean	SD					
Preliminary phase score	46.67	28.28	60.00	22.91	9	13.33	−2.51 to 29.18	0.52	0.088
Emergency phase score	52.12	12.79	54.02	12.26	9	1.90	−10.06 to 13.86	0.15	0.724
Management phase score	28.39	12.38	56.96	13.39	9	28.57	13.84 to 43.30	2.22	0.002
Total scenario score	41.44	8.77	55.70	9.56	9	14.26	4.71 to 23.82	1.56	0.009

*BCa 95% CIs are based on 5000 bootstrap samples.
BCa, bias-corrected and accelerated bootstrap.

leader inclusiveness (r_s = 0.392, P = 0.006), supporting hypothesis 5 (Table 3).

DISCUSSION

In this report, we aimed to provide a scenario development process facilitating specific targeting of learning objectives and their measurement during SBT. The resulting tool, PARTS, allows for separating complex cases into clearly delineated phases with single critical events. In addition, we tested PARTS for reliability and validity. With respect to reliability, we found that the PARTS scenario flowchart allowed for real-time rating providing similar results to postsimulation video-based rating. Concerning validity, we found that PARTS allowed for detecting differences in performance after an educational intervention as well as providing sensitivity in detecting associations among performance and leadership behavior.

Our findings suggest that PARTS may offer increased opportunity for assessment by phase augmentation. By identifying and isolating phase-specific critical events and required learner actions from real critical incidents and applying a visual technique resulting in a rating tool, PARTS contributed to reliable real-time rating and valid performance assessment. Extending EBAT scenario design around 1 critical event requiring learner actions, PARTS supports designers to chain critical events as encountered in complex cases, each with their respective required learner actions, while clearly separating these for individual rating.

LIMITATIONS

Although based on and integrating existing techniques for scenario design,[11–21] PARTS is a new instrument and has yet to prove its effectiveness beyond our reliability and validity tests, which we conducted in only 1 center.

Strengths and weaknesses of evaluation tools should be considered in light of the required assessment. For example, PARTS uses checklists for rating purposes. Among the weaknesses of checklists are the difficulty to rate unobservable events and the possible penalty to expert clinicians more apt to legitimately skip steps.[24] They can, however, produce reliable data,[64] achievable by novice and expert raters alike.[70] Because our aim was to create a formative rating instrument that could easily and reliably be used in real time by raters knowledgeable of the medical specialty without further training, we decided to use simple checklist scoring. We did, however, attempt to address their weaknesses (the difficulty of rating unobservable actions and the penalty to experts legitimately skipping steps) in our design process. Omitting tasks that cannot be observed from this list may affect the rating's validity and distort the performance measurement, but in our case, we found feasible surrogate actions in each case. Although selecting surrogate actions with less specificity for the required action may be acceptable in formative assessment, more rigorous standards should be applied to the rating, should this be required for summative assessment or competency assessment and certification.

Although the generated rating tool possesses interrater reliability and is suitable for real-time rating to provide formative feedback useful for debriefings, we have not yet extensively tested tool reliability, for example, by comparing various on-scene ratings, reliability over more simulation sessions, or results from additional simulation centers. In addition, the validity of the rating tool would have to be further established to compare our scores to global ratings.[24,36,51,76,77]

Furthermore, measurements presented here are subject to various biases. For a start, the same attending participated in all cases of the same day, and observation of previous cases will affect performance of groups participating later in 1 day. Nevertheless, we feel that total and phase-specific scores were affected in a similar direction, and hence, these results indicate the advantages of phase-specific measurement. In addition, the same raters performed postsimulation, video-based rating and counted the occurrence of leader inclusive behavior. Here, we hoped to reduce the common method bias by clearly defining the statements to be rated. In addition, the rater simultaneously controlling the scenario might be distracted from rating at certain times, such as when technical problems with software arise.

The PARTS design process likely is more time consuming than other methods and might be inappropriate for shorter or less complex scenarios, although the benefit of phase separation, selective focus, and simple rating should be balanced against effort. Moreover, subject to a common challenge in simulation, PARTS invariably fails to depict

TABLE 3. Spearman Rank-Order Correlation of Leader Inclusiveness With Standardized Performance Scores for Scenarios 1, 2, and 6 (n = 48 Scenarios)

Outcome	r_s	n	BCa 95% CI*	*P*
Preliminary phase score	0.001	48	−0.290 to 0.296	0.996
Emergency phase score	.392	48	0.118 to 0.632	0.006
Management phase score	−0.015	48	−0.334 to 0.302	0.917
Total scenario score	0.228	48	−0.082 to 0.520	0.119

*BCa 95% CIs are based on 5000 bootstrap samples.
BCa, bias-corrected and accelerated bootstrap.

critical incidents in their entirety by reducing complexity to improve ratings.

FUTURE RESEARCH

Further research is required to examine the benefits and suitability of PARTS and phase-specific rating across different samples and settings and to compare it with other assessment methods. In addition, we have yet to test PARTS in multidisciplinary training, but they seem well suited because phase transitions come naturally where specialist groups working in parallel might individually reach a common or individual interim achievement, leading on to the next phase. Structure, ratability, and discriminative properties of PARTS could be retained to facilitate assessment of individual disciplines or multidisciplinary teams alike.

CONCLUSIONS

Because measurements derived from SBT are important for research, program evaluation, and the substantiation of debriefing with formative assessment, we consider PARTS a valuable contribution for educators focusing their resources on high-standard simulation-based clinical education because they increase the opportunity for empirical measurement in realistic and complex cases.

REFERENCES

1. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003–2009. *Med Educ* 2010;44:50–63.
2. Fanning RM, Gaba DM. The role of debriefing in simulation-based learning. *Simul Healthc* 2007;2:115–125.
3. Rudolph JW, Simon R, Raemer DB, Eppich WJ. Debriefing as formative assessment: closing performance gaps in medical education. *Acad Emerg Med* 2008;15:1010–1016.
4. Levine AI, Flynn BC, Bryson EO, Demaria S Jr. Simulation-based Maintenance of Certification in Anesthesiology (MOCA) course optimization: use of multi-modality educational activities. *J Clin Anesth* 2012;24:68–74.
5. Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based objective structured clinical examination into the Israeli National Board Examination in Anesthesiology. *Anesth Analg* 2006;102:853–858.
6. Weller J, Morris R, Watterson L, et al. Effective management of anaesthetic crises: development and evaluation of a college-accredited simulation-based course for anaesthesia education in Australia and New Zealand. *Simul Healthc* 2006;1:209–214.
7. Hatala R, Kassen BO, Nishikawa J, Cole G, Issenberg SB. Incorporating simulation technology in a Canadian internal medicine specialty examination: a descriptive report. *Acad Med* 2005;80:554–556.
8. Gallagher AG, Cates CU. Approval of virtual reality training for carotid stenting: what this means for procedural-based medicine. *JAMA* 2004;292:3024–3026.
9. McGaghie WC, Draycott TJ, Dunn WF, Lopez CM, Stefanidis D. Evaluating the impact of simulation on translational patient outcomes. *Simul Healthc* 2011;6(suppl):S42–S47.
10. Woolf SH. The meaning of translational research and why it matters. *JAMA* 2008;299:211–213.
11. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology* 1998;89:8–18.
12. Künzle B, Zala-Mezö E, Kolbe M, Wacker J, Grote G. Substitutes for leadership in anaesthesia teams and their impact on leadership effectiveness. *European Journal of Work and Organizational Psychology* 2010;19:505–531.
13. Tschan F, Semmer NK, Hunziker PR, Marsch SCU. Decisive action vs. joint deliberation: different medical tasks imply different coordination requirements. *Advances in Human Factors and Ergonomics in Healthcare* 2011:191–200.
14. Tschan F, Semmer NK, Vetterli M, Gurtner A, Hunziker S, Marsch SU. Developing observational categories for group process research based on task and coordination requirement analysis: examples from research on medical emergency-driven teams. In: *Coordination in Human and Primate Groups*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011:93–115.
15. Clayton MJ. Delphi: a technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology* 1997;17:373–386.
16. Dieckmann P, Lippert A, Glavin R, Rall M. When things do not go as expected: scenario life savers. *Simul Healthc* 2010;5:219–225.
17. Rosen MA, Salas E, Wu TS, et al. Promoting teamwork: an event-based approach to simulation-based teamwork training for emergency medicine residents. *Acad Emerg Med* 2008;15:1190–1198.
18. Rosen MA, Salas E, Silvestri S, Wu TS, Lazzara EH. A measurement tool for simulation-based training in emergency medicine: the simulation module for assessment of resident targeted event responses (SMARTER) approach. *Simul Healthc* 2008;3:170–179.
19. Fowlkes JE, Lane NE, Salas E, Franz T, Oser R. Improving the measurement of team performance: the TARGETs methodology. *Military Psychology* 1994;6:47–61.
20. Fowlkes JE, Dwyer DJ, Oser RL, Salas E. Event-based approach to training (EBAT). *Int J Aviat Psychol* 1998;8:209–221.
21. Weingart LR. How did they do that? The ways and means of studying group processes. *Research in Organizational Behavior* 1997;19:189–240.
22. Kolbe M, Weiss M, Grote G, et al. TeamGAINS: a tool for structured debriefings for simulation-based team trainings. *BMJ Qual Saf* 2013;22:541–553.
23. Nembhard IM, Edmondson AC. Making it safe: the effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. *Journal of Organizational Behavior* 2006;27:941–966.
24. Adler MD, Vozenilek JA, Trainor JL, et al. Comparison of checklist and anchored global rating instruments for performance rating of simulated pediatric emergencies. *Simul Healthc* 2011;6:18–24.
25. Ahmed M, Sevdalis N, Paige J, Paragi-Gururaja R, Nestel D, Arora S. Identifying best practice guidelines for debriefing in surgery: a tri-continental study. *Am J Surg* 2012;203:523–529.
26. Beaubien M, Baker DP. The use of simulation for training teamwork skills in health care: how low can you go? *Qual Saf Health Care* 2004;13:i51–i56.
27. Birsner ML, Satin AJ. Developing a program, a curriculum, a scenario. *Semin Perinatol* 2013;37:175–178.
28. Blum RH, Boulet JR, Cooper JB, Muret-Wagstaff SL, ; Harvard Assessment of Anesthesia Resident Performance Research Group. Simulation-based assessment to identify critical gaps in safe anesthesia resident performance. *Anesthesiology* 2014;120:129–141.
29. Boulet JR, Murray D, Kras J, Woodhouse J. Setting performance standards for mannequin-based acute-care scenarios: an examinee-centered approach. *Simul Healthc* 2008;3:72–81.
30. Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology: requirements for practical implementation. *Anesthesiology* 2010;112:1041–1052.
31. Cantillon P, Wood D, Hutchinson L. ABC of learning and teaching in medicine. London: BMJ Books; 2003.
32. Carraccio CL, Benson BJ, Nixon LJ, Derstine PL. From the educational bench to the clinical bedside: translating the Dreyfus developmental model to the learning of clinical skills. *Acad Med* 2008;83:761–767.
33. Cristancho SM, Moussa F, Dubrowski A. A framework-based approach to designing simulation-augmented surgical education and training programs. *Am J Surg* 2011;202:344–351.
34. Cheng A, Rodgers DL, van der Jagt É, Eppich W, O'Donnell J. Evolution of the Pediatric Advanced Life Support course: enhanced learning with

a new debriefing tool and Web-based module for Pediatric Advanced Life Support instructors. *Pediatr Crit Care Med* 2012;13:589–595.

35. Cooper JB, Singer SJ, Hayes J, et al. Design and evaluation of simulation scenarios for a program introducing patient safety, teamwork, safety leadership, and simulation to healthcare leaders and managers. *Simul Healthc* 2011;6:231–238.
36. Donoghue A, Nishisaki A, Sutton R, Hales R, Boulet J. Reliability and validity of a scoring instrument for clinical performance during Pediatric Advanced Life Support simulation scenarios. *Resuscitation* 2010;81:331–336.
37. Donoghue A, Ventre K, Boulet J, et al. Design, implementation, and psychometric analysis of a scoring instrument for simulated pediatric resuscitation: a report from the EXPRESS pediatric investigators. *Simul Healthc* 2011;6:71–77.
38. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ* 2004;38:327–333.
39. Edler AA, Fanning RG, Chen MI, et al. Patient simulation: a literary synthesis of assessment tools in anesthesiology. *J Educ Eval Health Prof* 2009;6:3.
40. Fehr JJ, Boulet JR, Waldrop WB, Snider R, Brockel M, Murray DJ. Simulation-based assessment of pediatric anesthesia skills. *Anesthesiology* 2011;115:1308–1315.
41. Frost EA. *Comprehensive Guide to Education in Aanesthesia*. Springer; 2014.
42. Gaba DM, DeAnda A. A comprehensive anesthesia simulation environment: re-creating the operating room for research and training. *Anesthesiology* 1988;69:387–394.
43. Gaba DM. The future vision of simulation in health care. *Qual Saf Health Care* 2004;13:i2–i10.
44. Gerard JM, Kessler DO, Braun C, Mehta R, Scalzo AJ, Auerbach M. Validation of global rating scale and checklist instruments for the infant lumbar puncture procedure. *Simul Healthc* 2013;8:148–154.
45. Gordon JA, Tancredi DN, Binder WD, Wilkerson WM, Shaffer DW. Assessment of a clinical performance evaluation tool for use in a simulator-based testing environment: a pilot study. *Acad Med* 2003;78:S45–S47.
46. Gordon JA, Oriol NE, Cooper JB. Bringing good teaching cases "to life": a simulator-based medical education service. *Acad Med* 2004;79:23–27.
47. Issenberg SB, McGaghie WC, Hart IR, et al. Simulation technology for health care professional skills training and assessment. *JAMA* 1999;282:861–866.
48. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005;27:10–28.
49. Issenberg SB, Ringsted C, Ostergaard D, Dieckmann P. Setting a research agenda for simulation-based healthcare education: a synthesis of the outcome from an Utstein style meeting. *Simul Healthc* 2011;6:155–167.
50. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Crit Care Med* 2006;34:2167–2174.
51. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J. High-fidelity patient simulation: validation of performance checklists. *Br J Anaesth* 2004;92:388–392.
52. Morgan PJ, Lam-McCulloch J, Herold-McIlroy J, Tarshis J. Simulation performance checklist generation using the Delphi technique. *Can J Anaesth* 2007;54:992–997.
53. Morgan PJ, Cleave-Hogg D, Guest CB. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. *Acad Med* 2001;76:1053–1055.
54. Morgan PJ, Cleave-Hogg DM, Guest CB, Herold J. Validity and reliability of undergraduate performance assessments in an anesthesia simulator. *Can J Anaesth* 2001;48:225–233.
55. Morgan PJ, Kurrek MM, Bertram S, LeBlanc V, Przybyszewski T. Nontechnical skills assessment after simulation-based continuing medical education. *Simul Healthc* 2011;6:255–259.
56. Neal JM, Hsiung RL, Mulroy MF, Halpern BB, Dragnich AD, Slee AE. ASRA checklist improves trainee performance during a simulated episode of local anesthetic systemic toxicity. *Reg Anesth Pain Med* 2012;37:8–15.
57. Norcini J. Being smarter about SMARTER: a commentary on "a measurement tool for simulation-based training in emergency medicine: the simulation module for assessment of resident targeted event responses approach". *Simul Healthc* 2008;3:131–132.
58. Pastis NJ, Doelken P, Vanderbilt AA, Walker J, Schaefer JJ 3rd. Validation of simulated difficult bag-mask ventilation as a training and evaluation method for first-year internal medicine house staff. *Simul Healthc* 2013;8:20–24.
59. Pierre MS, Breuer G. *Simulation in der medizin: Grundlegende konzepte–klinische anwendung*. Springer-Verlag; 2013.
60. Raemer D, Anderson M, Cheng A, Fanning R, Nadkarni V, Savoldelli G. Research regarding debriefing as part of the learning process. *Simul Healthc* 2011;6(suppl):S52–S57.
61. Reznek M, Smith-Coggins R, Howard S, et al. Emergency medicine crisis resource management (EMCRM): pilot study of a simulation-based crisis management course for emergency medicine. *Acad Emerg Med* 2003;10:386–389.
62. Riley RH. *Manual of Simulation in Healthcare*. Oxford, New York: Oxford University Press; 2008.
63. Rudolph JW, Simon R, Dufresne RL, Raemer DB. There's no such thing as "nonjudgmental" debriefing: a theory and method for debriefing with good judgment. *Simul Healthc* 2006;1:49–55.
64. Salas E, Rosen A, Held D, Weissmuller J. Performance measurement in simulation-based training: a review and best practices. *Simulation & Gaming* 2009;40:328–376.
65. Scalese RJ, Obeso VT, Issenberg SB. Simulation technology for skills training and competency assessment in medical education. *J Gen Intern Med* 2008;23(suppl 1):46–49.
66. Schwid HA. Anesthesia simulators—technology and applications. *Isr Med Assoc J* 2000;2:949–953.
67. Schwid HA, Rooke GA, Carline J, et al. Evaluation of anesthesia residents using mannequin-based simulation: a multiinstitutional study. *Anesthesiology* 2002;97:1434–1444.
68. Seropian MA. General concepts in full scale simulation: getting started. *Anesth Analg* 2003;97:1695–1705.
69. Sinz EH. Anesthesiology national CME program and ASA activities in simulation. *Anesthesiol Clin* 2007;25:209–223.
70. Stout RJ, Salas E, Fowlkes JE. Enhancing teamwork in complex environments through team training. *Group Dyn* 1997;1:169–182.
71. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;357:945–949.
72. Ziv A, Rubin O, Sidi A, Berkenstadt H. Credentialing and certifying with simulation. *Anesthesiol Clin* 2007;25:261–269.
73. Künzle B, Zala-Mezö E, Wacker J, Kolbe M, Spahn DR, Grote G. Leadership in anaesthesia teams: the most effective leadership is shared. *Qual Saf Health Care* 2010;19:e46.
74. Rudolph JW, Morrison JB, Carroll JS. The dynamics of action-oriented problem solving: linking interpretation and choice. *Acad Manage Rev* 2009;34:733–756.
75. Van Heukelom JN, Begaz T, Treat R. Comparison of postsimulation debriefing versus in-simulation debriefing in medical simulation. *Simul Healthc* 2010;5:91–97.
76. Mudumbai SC, Gaba DM, Boulet JR, Howard SK, Davies MF. External validation of simulation-based assessments with other performance measures of third-year anesthesiology residents. *Simul Healthc* 2012;7:73–80.
77. Swartz MH, Colliver JA, Bardes CL, Charon R, Fried ED, Moroff S. Validating the standardized-patient assessment administered to medical students in the New York City Consortium. *Acad Med* 1997;72:619–626.

